控制与决策
Control and Decision

基于代价敏感的粗糙集近似集与粒度寻优算法

张清华， 刘凯旋， 高满

引用本文:
张清华, 刘凯旋, 高满. 基于代价敏感的粗糙集近似集与粒度寻优算法[J]. 控制与决策, 2020, 35(9): 2070−2080.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.0149

基于代价敏感的粗糙集近似集与粒度寻优算法

张清华†, 刘凯旋, 高 满

(1. 重庆邮电大学 计算机科学与技术学院，重庆 400065；
2. 重庆邮电大学 计算智能重庆市重点实验室，重庆 400065)

摘 要: 粗糙集的近似集用已有知识粒对不确定性目标概念进行近似描述,但在构建近似集时并没有考虑数据的代价信息这一实际因素. 对此,首先分析在构建粗糙集的近似集时考虑代价信息的必要性;然后,从代价敏感角度构建误分类代价的粗糙集近似集模型,并分析该模型下求得的近似集的相关性质. 为了在多粒度空间中寻找一个合适的粒度空间来对不确定性目标概念进行近似描述,使误分类代价与测试代价之和尽可能小,给出属性代价贡献率的定义,并提出一种代价敏感的粒度寻优算法. 实验结果表明,所提出算法能适用于现有代价认知场景,并在给定代价场景下求出合理的层次粒度空间结构以及不确定性目标概念的近似集.

关键词: 粗糙集；近似集；代价敏感；多粒度；粒度寻优

中图分类号: TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0149

引用格式: 张清华,刘凯旋,高满. 基于代价敏感的粗糙集近似集与粒度寻优算法[J]. 控制与决策, 2020, 35(9): 2070-2080.

Approximation sets of rough sets and granularity optimization algorithm based on cost-sensitive

ZHANG Qing-hua†, LIU Kai-xuan, GAO Man

(1. School of Computer Science and Technology，Chongqing University of Posts and Telecommunications，Chongqing 400065，China；2. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications，Chongqing 400065，China)

Abstract: Approximation sets of rough sets use existing knowledge granules to describe the uncertain concept approximately. However, the cost information contained in data have not been considered when constructing the approximation set of the uncertain concept. Therefore, the necessity of considering cost information is analyzed when constructing the approximation sets of rough sets firstly. Then an approximation set model of rough sets of misclassification cost is constructed from the perspective of cost-sensitive, and some properties of this model are discussed in detail. In order to find a suitable granularity space to describe the uncertain concept approximately that can minimize the sum of misclassification cost and test cost as far as possible in multi-granulation spaces, the contribute rate of attribute cost is defined, and the cost-sensitive granularity optimization algorithm is proposed. The experimental results show that the proposed algorithm can be used in existing cost cognition scene. And a reasonable hierarchy granulation space and the approximation set of the uncertain concept can be obtained under the given cost scene.

Keywords: rough sets；approximation sets；cost-sensitive；multi-granularity；granularity optimization

0 引 言

目前,随着计算机技术以及认知科学的发展,越来越多的专家学者开始聚焦于不确定性问题的研究. 自Zadeh[1]提出模糊集理论以来,不确定性问题的研究取得了重大突破,各种处理不确定性问题的理论也随之产生,例如,Pawlak[2-3]提出的粗糙集理论,张钹等[4-5]提出的商空间理论,李德毅等[6-7]提出的云模型理论等. 作为处理不确定性问题的重要方法,这些理论已被广泛地用于人工智能领域. 粗糙集理论在处理不确定性问题时,无需提供所需处理数据集合之外的任何先验知识,它与模糊集、概率论、证据理论等其他理论有着很强的互补性. 因此,粗糙集已成为

收稿日期: 2019-02-01；**修回日期:** 2019-04-23.
基金项目: 国家自然科学基金项目(61876201)；重庆市研究生科研创新项目(CYS18244).
责任编委: 薛建儒.
†通讯作者. E-mail: zhangqh@cqupt.edu.cn.

一种重要的智能信息处理技术[8-12],越来越受到广大专家学者的青睐.

为了使粗糙集理论能更好地应用于问题求解中,许多专家学者从不同角度构建了粗糙集的扩展模型,如概率粗糙集[13]、模糊粗糙集[14]、决策粗糙集[15-16]、变精度粗糙集[17]、邻域粗糙集[18-19]等. 这些模型构建了扩展的Pawlak近似算子,丰富了粗糙集理论,但是没有考虑用现有知识粒构建目标概念的近似集[20]. 为了能用现有知识粒构建一个精确集对不精确集进行近似刻画,在前期的研究中,笔者从集合相似度出发,利用模糊截集提出了粗糙集的近似集模型[20-21],并且基于该模型在属性约简、图像分割、文本分类等领域取得了一系列成果[22-24],这些研究扩展了粗糙集理论模型及应用.

现实中,许多数据都含有代价信息,将代价信息引入到数据挖掘及机器学习领域中便产生了代价敏感学习方法[25-27]. 许多专家学者将代价敏感学习运用于粗糙集理论中,并取得了重要的研究成果[28-33]. 在代价敏感学习中考虑最多的两种代价为测试代价和误分类代价[34]. 在前期的粗糙集近似集模型研究中,实现了利用现有知识粒来构建目标概念的近似集,但是,此时构成的近似集必然会存在对象错误划分的情况. 由于人类在对现实问题决策时,往往希望得到的划分结果的误分类代价尽可能小,考虑到代价这一实际因素,从集合相似度出发构建的近似集很有可能不适应现实应用的需求. 因此,为使所构建的近似集符合人类实际决策,基于代价敏感角度,本文构建一种误分类代价的粗糙集近似集模型,以保证得到的近似集的误分类代价总和在当前知识空间下最小,该模型也更加符合人类认知,在现实场景中具有更好的应用价值.

现实中,人们往往希望问题求解的精度尽可能高,并且所产生的误分类代价尽可能低. 为了有效降低构建近似集所产生的误分类代价,人们会考虑引入新的属性来细化当前粒度空间. 然而,现实中属性值的获取需要付出一定的测试代价. 尽管随着粒度空间的细化,所求得的近似集的误分类代价会逐渐降低,但也会造成测试代价的上升,而且在不同粒度空间下,人们的决策结果往往是不同的. 为了使在多粒度空间中求得的近似集更加符合人类认知,出于代价优化的考虑,应将误分类代价与测试代价综合考虑. 测试代价的产生主要是由于边界域对象的不确定性,通常,人们为了更精确地对边界域中的对象进行决策划分,会进一步获取边界域对象的新的属性值,而引入不同的属性所带来的测试代价是不一样的. 为了选取合理的属性细化已知粒度空间,本文给出属性代价贡献率的定义,并根据该定义提出代价敏感的粒度寻优算法. 该算法可以在不同粒度空间下有效选取合适的属性来细化当前粒度空间,进而在一个合理的粒度空间下构建目标概念的近似集,帮助人们从数据中挖掘出具有代价敏感的决策知识. 最后,通过实验表明,对于同一数据集以及同一目标概念,在不同的代价场景下将产生不同的层次空间结构,所得到的近似集也更加符合实际应用场景,这也是代价敏感的充分体现.

1　相关基本概念

为了能更清楚地对本文进行阐述,这里首先给出相关的基本概念.

定义1 (决策信息系统)[2]　一个决策信息系统S可以表示为

$$S=(U,A,V,f).$$

其中:U是对象全集,也称为论域;$A=C\bigcup D$是属性全集,C和D分别为条件属性集和决策属性集;$V=\bigcup\limits_{r\in A}V_r$是属性值的集合,$V_r$表示属性$r\in A$的属性值范围,即属性$r$的值域;$f:U\times A\to V$是一个信息函数,它指定$U$中每一个对象$x$的属性值.

定义2 (粗糙集)[2]　给定决策信息系统$S=(U,A,V,f)$,对于任一对象集合$X\subseteq U$和属性集合$R\subseteq C$,当X能用属性子集R确切地描述时,称X是R可定义的. R可定义集称为R精确集,R不可定义集称为R不精确集或R粗糙集. 当X是R粗糙集时,X的R上近似集和R下近似集分别为

$$\overline{R}(X)=\bigcup\{Y_i|Y_i\in U/\mathrm{IND}(R)\bigwedge Y_i\bigcap X\neq\varnothing\},\tag{1}$$

$$\underline{R}(X)=\bigcup\{Y_i|Y_i\in U/\mathrm{IND}(R)\bigwedge Y_i\subseteq X\}.\tag{2}$$

集合$\mathrm{BND}_R(X)=\overline{R}(X)-\underline{R}(X)$称为$X$的$R$边界域,$\mathrm{POS}_R(X)=\underline{R}(X)$称为$X$的$R$正域,$\mathrm{NEG}_R(X)=U-\overline{R}(X)$称为$X$的$R$负域. 其中:$\mathrm{IND}(R)$为一个不分明关系(等价关系),$\mathrm{IND}(R)=\{(x,y)|(x,y)\in U^2,\ \forall\limits_{r\in R}(r(x)=r(y))\}$. $U/\mathrm{IND}(R)=\{E|(E\subseteq U\bigwedge\forall\limits_{x\in E,y\in E,r\in R}(r(x)=r(y)))\}$是不分明关系$\mathrm{IND}(R)$在$U$上的划分.

定义3 (粗糙集的近似集)[20]　给定决策信息系统$S=(U,A,V,f)$,对于任意不确定性目标概念$X(X\subseteq U)$和属性子集$R(R\subseteq A)$,令

$$R_\alpha(X)=\{x\in U|\mu_X^R(x)\geqslant\alpha\},\ 0<\alpha\leqslant 1,\tag{3}$$

称$R_\alpha(X)$为X的α近似集. 其中:在$\mathrm{IND}(R)$的划分下,$\mu_X^R(x)$表示对于任意的$x(x\in U)$,x属于集合

X 的隶属程度,$\mu_X^R(x) = \dfrac{|X \bigcap [x]_R|}{|[x]_R|}$. 这里:$[x]_R \in U/\mathrm{IND}(R)$,$[x]_R$ 表示对象 x 在不分明关系 $\mathrm{IND}(R)$ 上形成的等价类(划分块).

2 基于误分类代价的粗糙集近似集

粗糙集给出了不确定性目标概念的两个边界线,但它无法给出一个不确定性概念外延的近似描述. 粗糙集的 α 近似集模型实现了用已有知识粒构建目标概念的近似集,通过对比 α 与对象 x 隶属于目标概念 X 的隶属程度来判定该对象是否可用于对 X 作近似描述. 如图1所示,位于边界域中的等价类 E_1 中有6个对象,其中4个圆形对象属于目标概念 X,2个三角形对象不属于 X. 假设要求构建 X 的近似集 $R_{0.5}(X)$,根据 E_1 隶属于 X 的隶属度,可知 $E_1 \subset R_{0.5}(X)$,即 E_1 将被用于对 X 进行近似描述.

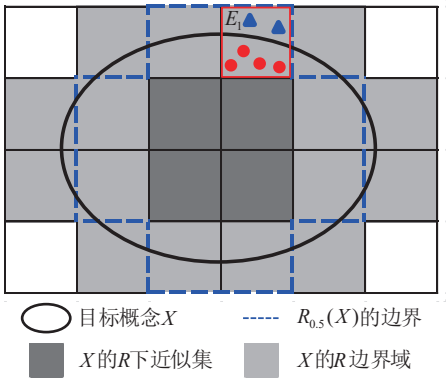


图1 目标概念的 $R_{0.5}(X)$ 近似集

现实中,如果事物出现错误划分,则必然会因错误划分而付出一定的误分类代价. 在前期研究中,构造目标概念的近似集时并没有考虑误分类代价这一现实因素. 考虑这一因素,本文给出如下误分类代价的决策信息系统.

定义4(误分类代价的决策信息系统) 一个误分类代价的决策信息系统 S 可以表示为 $S = (U, A, V, f, \lambda^e)$. 其中:$U$、$A$、$V$、$f$ 的定义与定义1中的一致,$\lambda^e = \{\lambda_{x_1}, \lambda_{x_2}, \ldots, \lambda_{x_{|U|}}\}$,$\lambda_{x_i}(1 \leqslant i \leqslant |U|)$ 表示 U 中对象 x_i 的误分类代价取值.

虽然在给定粒度空间下,同一等价类中各个对象的属性值一样,但随着粒度空间的不断细化,每一个对象都会被区分开来,因此,对于论域中的每一个对象,它们的误分类代价不一定相同. 对于对象 $x_i \in U$,将其误分类代价记为 λ_{x_i},如果 $x_i \in X$,则当它被用来描述 X 时,$\lambda_{x_i} = 0$;如果 $x_i \notin X$,则当它不被用来描述 X 时,$\lambda_{x_i} = 0$. 等价类 E_j 的误分类代价记为 λ_{E_j}.

在粗糙集中,可以清晰地判断正域和负域中的等价类是否属于目标概念,因此,在构建 $R_{0.5}(X)$ 时,正域和负域中的等价类不产生误分类代价. 而对于某一等价块 $E_j \subseteq \mathrm{BND}_R(X)$,$E_j$ 无论是否被用来描述目标概念,总会产生误分类代价. 假设在图1中,2个三角形对象的误分类代价之和为 λ_1,4个圆形对象的误分类代价之和为 λ_2,当 $\lambda_1 > \lambda_2$ 时,如果继续将 E_1 用于对 X 进行近似描述,则显然是不合理的. 因此,结合实际生产需求,构建误分类代价的粗糙集近似集模型是非常必要的. 对于等价类 E_j,当 E_j 被用于描述 X 时,E_j 的误分类代价为

$$\lambda_{E_j}^Y = \sum_{x_i \in E_j \wedge x_i \in \sim X} \lambda_{x_i}; \tag{4}$$

当 E_j 不被用于描述 X 时,E_j 的误分类代价为

$$\lambda_{E_j}^N = \sum_{x_i \in E_j \wedge x_i \in X} \lambda_{x_i}. \tag{5}$$

从代价最小化角度出发,如果 $\lambda_{E_j}^Y \leqslant \lambda_{E_j}^N$,则 E_j 将被用于对 X 进行近似描述;否则,E_j 将不被用于对 X 进行近似描述. 基于误分类代价的决策信息系统,本文给出误分类代价的粗糙集近似集的定义.

定义5(误分类代价的粗糙集近似集) 给定误分类代价的决策信息系统 $S = (U, A, V, f, \lambda^e)$,设 X 是论域上的一个集合(目标概念),$R \subseteq C$,令

$$R(X) = \bigcup \{E_i \subseteq U | \lambda_{E_i}^Y \leqslant \lambda_{E_i}^N\},$$

其中 $E_i \in U/\mathrm{IND}(R)$. 称 $R(X)$ 为误分类代价敏感下 X 的近似集. 用 $\lambda_{R(X)}$ 表示 $R(X)$ 作为 X 的近似集而产生的误分类代价,即 $\lambda_{R(X)} = \sum_{E_i \in U/\mathrm{IND}(R)} \min(\lambda_{E_i}^Y, \lambda_{E_i}^N)$,显然 $\underline{R}(X) \subseteq R(X) \subseteq \overline{R}(X)$.

为了更清晰地阐述误分类代价的粗糙集近似集模型,下面给出一个例子.

例1 给定如表1所示的误分类代价的决策信息系统 $S = (U, A, V, f, \lambda^e)$. 其中:$U = \{x_1, x_2, \ldots, x_{10}\}$,$X = \{x_1, x_3, x_4, x_7, x_8, x_{10}\}$,$R = \{a_1, a_2, a_3\}$. 可知 $U/\mathrm{IND}(R) = \{E_1, E_2, E_3, E_4\}$. 其中:$E_1 = \{x_1\}$,$E_2 = \{x_2, x_3, x_4\}$,$E_3 = \{x_5, x_6, x_7, x_8\}$,$E_4 = \{x_9, x_{10}\}$. 根据对象中的误分类代价计算得 $\lambda_{E_1}^Y < \lambda_{E_1}^N$,$\lambda_{E_2}^Y > \lambda_{E_2}^N$,$\lambda_{E_3}^Y < \lambda_{E_3}^N$,$\lambda_{E_4}^Y > \lambda_{E_4}^N$,根据定义5可得 $R(X) = E_1 \bigcup E_3$,即 $R(X) = \{x_1, x_5, x_6, x_7, x_8\}$.

表1 误分类代价的决策信息系统

U	a_1	a_2	a_3	a_4	d	λ^e
x_1	0	0	0	0	1	5
x_2	1	0	1	2	0	4
x_3	1	0	1	2	1	1
x_4	1	0	1	0	1	2
x_5	0	1	1	1	0	1
x_6	0	1	1	1	0	2
x_7	0	1	1	2	1	3
x_8	0	1	1	2	1	2
x_9	1	1	0	1	0	5
x_{10}	1	1	0	1	1	3

下面详细讨论$R(X)$的一些相关运算性质.

性质1　给定误分类代价的决策信息系统$S=(U,A,V,f,\lambda^e)$,设X、Y是论域上的两个集合,$\sim X$表示X的补集,$R\subseteq C$,则:

1) $R(\sim X)=\sim R(X)$;

2) 若$X\subseteq Y$,则$R(X)\subseteq R(Y)$;

3) $R(X\bigcup Y)\supseteq R(X)\bigcup R(Y)$;

4) $R(X\bigcap Y)\subseteq R(X)\bigcap R(Y)$;

5) $\sim(R(X)\bigcup R(Y))=R(\sim X)\bigcap R(\sim Y)$;

6) $\sim(R(X)\bigcap R(Y))=R(\sim X)\bigcup R(\sim Y)$.

证明　1) 对于$\forall E_i\in U/\mathrm{IND}(R)$,如果$E_i\subseteq R(X)$,则$E_i\not\subset R(\sim X)$,可知$R(\sim X)\bigcup R(X)=U$,即$R(\sim X)=\sim R(X)$,得证.

2) 对于$\forall E_i\in U/\mathrm{IND}(R)$,因为$X\subseteq Y$,所以$\overline{R}(X)\subseteq\overline{R}(Y)$. 当$\overline{R}(X)=\overline{R}(Y)$时,如果$E_i\subseteq R(X)$($E_i\not\subset R(X)$),则对于目标概念$X$而言,$\lambda^Y_{E_i}\leqslant\lambda^N_{E_i}$($\lambda^Y_{E_i}>\lambda^N_{E_i}$);同样,对于目标概念$Y$而言,显然$\lambda^Y_{E_i}\leqslant\lambda^N_{E_i}(\lambda^Y_{E_i}>\lambda^N_{E_i})$成立,此时$R(X)=R(Y)$. 当$\overline{R}(X)\subset\overline{R}(Y)$时,必然有$\underset{E_j\in U/\mathrm{IND}(R)}{\exists}(E_j\bigcap X=\varnothing\bigwedge E_j\bigcap Y\neq\varnothing)$,此时,若$\forall E_j$不用于对$Y$进行近似描述,则$R(X)=R(Y)$成立;若$\exists E_j$用于对$Y$进行近似描述,则$R(X)\subset R(Y)$. 因此$R(X)\subseteq R(Y)$,得证.

3) 因为$X\subseteq X\bigcup Y$,由2)可知$R(X)\subseteq R(X\bigcup Y)$;同样有$R(Y)\subseteq R(X\bigcup Y)$,因此$R(X\bigcup Y)\supseteq R(X)\bigcup R(Y)$显然成立.

4) 同3)可得,该性质显然成立.

5) 由德·摩根定律可知$\sim(R(X)\bigcup R(Y))=\sim R(X)\bigcap\sim R(Y)$,由1)可知$\sim R(X)\bigcap\sim R(Y)=R(\sim X)\bigcap R(\sim Y)$,故$\sim(R(X)\bigcup R(Y))=R(\sim X)\bigcap R(\sim Y)$,得证.

6) 同5)可得,该性质显然成立. □

在不同粒度空间下求得的目标概念的误分类代价的近似集不一定相同,下面通过定理1来讨论不同粒度空间下误分类代价的近似集的误分类代价的变化规律.

定理1　给定误分类代价的决策信息系统$S=(U,A,V,f,\lambda^e)$,令$X\subseteq U,R\subseteq P\subseteq C,\lambda_{R(X)}$、$\lambda_{P(X)}$分别表示$R(X)$与$P(X)$下产生的误分类代价,则$\lambda_{R(X)}\geqslant\lambda_{P(X)}$.

证明　假设$U/\mathrm{IND}(R)=\{E_1,E_2,\ldots,E_y\}$,$U/\mathrm{IND}(P)=\{F_1,F_2,\cdots,F_z\}$,由于$R\subseteq P\subseteq C$,可知$U/\mathrm{IND}(R)$中的任意一个等价类可由$U/\mathrm{IND}(P)$中的一个或多个等价类合并而成.

任取$E_i\subseteq U(1<i<y)$,当$E_i\subseteq\mathrm{POS}_R(X)$或$E_i\subseteq\mathrm{NEG}_R(X)$时,不产生误分类代价;当$E_i\subseteq\mathrm{BND}_R(X)$时,无论是否将$E_i$用于对$X$的近似描述,都将产生误分类代价. 假设$E_i=\{F_j,F_{j+1},\ldots,F_{j+k}\}(1\leqslant j+k\leqslant z)$,如果存在$F_{j+m}\subseteq\mathrm{POS}_P(X)(0\leqslant m\leqslant k)$或$F_{j+m}\subseteq\mathrm{NEG}_P(X)(0\leqslant m\leqslant k)$,$F_{j+m}$不用于对目标概念进行近似描述,此时$F_{j+m}$的误分类代价为0,则$\min(\lambda^Y_{F_j},\lambda^N_{F_j})+\min(\lambda^Y_{F_{j+1}},\lambda^N_{F_{j+1}})+\ldots+\min(\lambda^Y_{F_{j+k}},\lambda^N_{F_{j+k}})\leqslant\min(\lambda^Y_{E_i},\lambda^N_{E_i})$,此时$\lambda_{R(X)}\geqslant\lambda_{P(X)}$显然成立. 当$\underset{F_{j+m}\subseteq E_i}{\forall}(F_{j+m}\subseteq\mathrm{BND}_P(X))(0\leqslant m\leqslant k)$时,可知$\lambda^Y_{E_i}=\sum_{l=0}^{k}\lambda^Y_{F_{j+l}},\lambda^N_{E_i}=\sum_{l=0}^{k}\lambda^N_{F_{j+l}}$,显然$\sum_{l=0}^{k}\lambda^Y_{F_{j+l}}\geqslant\sum_{l=0}^{k}\min(\lambda^Y_{F_{j+l}},\lambda^N_{F_{j+l}})$且$\sum_{l=0}^{k}\lambda^N_{F_{j+l}}\geqslant\sum_{l=0}^{k}\min(\lambda^Y_{F_{j+l}},\lambda^N_{F_{j+l}})$,故$\min(\lambda^Y_{E_i},\lambda^N_{E_i})\geqslant\sum_{l=0}^{k}\min(\lambda^Y_{F_{j+l}},\lambda^N_{F_{j+l}})$,此时$\lambda_{R(X)}\geqslant\lambda_{P(X)}$. 综上所述,当$R\subseteq P\subseteq C$时,$\lambda_{R(X)}\geqslant\lambda_{P(X)}$. □

定理1表明,在较细粒度空间下求得X的误分类代价的近似集的误分类代价不大于较粗粒度空间下X的误分类代价的近似集的误分类代价,这也符合人类认知.

在问题求解中,可以用$R(X)$、$\underline{R}(X)$和$\overline{R}(X)$分别作为X的近似集,那么采用哪一个集合作为X的近似集将会产生最少的误分类代价?下面将通过定理2 说明$R(X)$作为X的近似集在误分类代价方面的优势.

定理2　给定误分类代价的决策信息系统$S=(U,A,V,f,\lambda^e)$,令$X\subseteq U,R\subseteq C,\lambda_{R(X)}$、$\lambda_{\underline{R}(X)}$和$\lambda_{\overline{R}(X)}$分别表示$R(X)$、$\underline{R}(X)$和$\overline{R}(X)$作为$X$的近似集时产生的误分类代价,则$\lambda_{R(X)}\leqslant\lambda_{\underline{R}(X)},\lambda_{R(X)}\leqslant\lambda_{\overline{R}(X)}$.

证明　当用$\underline{R}(X)$作为X的近似集时,误分类代价为$\lambda_{\underline{R}(X)}=\sum_{E_i\in\mathrm{BND}_R(X)}\lambda^N_{E_i}$;当用$\overline{R}(X)$作为$X$的近似集时,误分类代价为$\lambda_{\overline{R}(X)}=\sum_{E_i\in\mathrm{BND}_R(X)}\lambda^Y_{E_i}$;当用$R(X)$作为$X$的近似集时,误分类代价为$\lambda_{R(X)}=\sum_{E_i\in\mathrm{BND}_R(X)}\min(\lambda^Y_{E_i},\lambda^N_{E_i})$,显然

$$\sum_{E_i\in\mathrm{BND}_R(X)}\min(\lambda^Y_{E_i},\lambda^N_{E_i})\leqslant\sum_{E_i\in\mathrm{BND}_R(X)}\lambda^N_{E_i},$$

$$\sum_{E_i\in\mathrm{BND}_R(X)}\min(\lambda^Y_{E_i},\lambda^N_{E_i})\leqslant\sum_{E_i\in\mathrm{BND}_R(X)}\lambda^Y_{E_i},$$

即$\lambda_{R(X)}\leqslant\lambda_{\underline{R}(X)},\lambda_{R(X)}\leqslant\lambda_{\overline{R}(X)}$. □

定理3　给定误分类代价的决策信息系统$S=$

(U,A,V,f,λ^e),令$X\subseteq U,R\subseteq C$,如果$\underset{x_i,x_j\in U}{\forall}(\lambda_{x_i}=\lambda_{x_j})$,则$R(X)=R_{0.5}(X)$.

证明 当$\underset{x_i,x_j\in U}{\forall}(\lambda_{x_i}=\lambda_{x_j})$时,$\lambda_{E_i}^Y\leqslant\lambda_{E_i}^N$表明$|E_i-X|\leqslant|X\bigcap E_i|$,由此可知$|X\bigcap E_i|\geqslant|E_i|/2$,即$\dfrac{|X\bigcap E_i|}{|E_i|}\geqslant\dfrac{1}{2}$,所以$\underset{x\in E_i}{\forall}\mu_x^R(x)\geqslant 0.5$. 又因为$R_{0.5}(X)=\{x|\mu_X^R(x)\geqslant 0.5\}$,所以$R(X)=R_{0.5}(X)$. □

定理3表明,当论域中所有对象的误分类代价都相等时,X的误分类代价的近似集$R(X)$将退化为$R_{0.5}(X)$.

本节提出的误分类代价的粗糙集近似集模型中,可以在给定粒度空间中,求出误分类代价最小的近似集用于对目标概念进行近似描述,并且通过对该近似集的分析,表明了所求得的近似集在代价敏感环境下更符合当前粒度空间下的认知环境.

3 代价敏感的粒度寻优策略

在上节中本文利用现有知识粒构建了目标概念误分类代价的近似集模型. 为了进一步减小构建近似集的误分类代价,只有引入新的属性来细化原有知识空间,但这也必然会造成测试代价的增加. 现实问题求解中,存在以付出较少测试代价换取降低较大误分类代价的情况,从而达到降低总代价的目的. 但是,当粒度空间细化到一定程度时,只能以较大的测试代价的付出换取较小的误分类代价的降低,此时出于对总代价优化的考虑,再对粒度空间细化已经没有意义了. 因此,在构建多粒度空间下代价敏感的粗糙集近似集时,从代价优化的角度出发,在选取粒度空间时,需要在误分类代价与测试代价之间寻找一个平衡点,进而得到一个合适的粒度空间来对目标概念进行近似描述. 考虑到测试代价和误分类代价这两种代价信息,本文将引入Min等[35]提出的测试代价和误分类代价的决策信息系统,如定义6所示.

定义6 (测试代价和误分类代价的决策信息系统)[35] 一个测试代价和误分类代价的决策信息系统S可以表示为$S=(U,A,V,f,\lambda^e,\lambda^t)$. 其中:$U$、$A$、$V$、$f$、$\lambda^e$的定义与定义4中的一致,$\lambda^t=\{\lambda_{a_1}^t,\lambda_{a_2}^t,\ldots,\lambda_{a_{|A|}}^t\}$,$\lambda_{a_j}^t(0\leqslant j\leqslant|A|)$表示任意对象获取属性$a_j$属性值所需的测试代价.

在给定粒度空间下,只有边界域对象具有不确定性,因此,在对当前粒度空间细化时,只需要获取边界域对象的属性值,因而测试代价主要是由边界域元素引起的. 在给定粒度空间下,选取不同的属性,所花费的测试代价是不同的. 那么在当前粒度空间下,如何选取新的属性从而构建新的粒度空间是一个研究要点. 由于人类认知有限,往往不能精确地给出所需要的属性集以达到代价最优化的目的,人类只能依赖当前认知体系来细化已有的粒度空间. 在对问题进一步细化时,可以从多角度出发,即引入不同的属性或属性集. 假设当前的粒度空间由属性集R诱导得出,并且通过新增属性a_1、a_2可以达到问题求解的目的. 如果一次性加入属性集$\{a_1,a_2\}$,则产生的测试代价为$(\lambda_{a_1}^t+\lambda_{a_2}^t)\mathrm{BND}_R(X)$;如果逐步增加属性,即先增加属性$a_1$,再增加属性$a_2$,则产生的测试代价为$\lambda_{a_1}^t\mathrm{BND}_R(X)+\lambda_{a_2}^t\mathrm{BND}_{R\cup\{a_1\}}(X)$,显然$(\lambda_{a_1}^t+\lambda_{a_2}^t)\mathrm{BND}_R(X)\geqslant\lambda_{a_1}^t\mathrm{BND}_R(X)+\lambda_{a_2}^t\mathrm{BND}_{R\cup\{a_1\}}(X)$. 虽然新的粒度空间都是由属性集$R\bigcup\{a_1,a_2\}$诱导得出的,但是,所带来的测试代价却是不一样的. 通过逐步增加属性可以使测试代价渐近式增长,因此在引入新的属性时,应考虑每次只引入一个属性,逐步缩减边界域.

由于属性选取引起的边界域动态变化,不同的属性集下产生的近似集是不同的. 属性选取的先后顺序对所求得的粒度空间构建近似集的总代价起关键作用,因此,构建合理的属性选取顺序也是非常必要的. 现实生活中,人类在当前认知环境下处理问题时,总想以最小的付出赢取最大的利益,即以最小的测试代价尽可能最大地降低误分类代价. 基于这一思想,本文给出属性代价贡献率的定义来对不同粒度空间下的属性进行选取,以达到属性的合理选取的目的.

定义7 (属性代价贡献率) 给定测试代价和误分类代价的决策信息系统$S=(U,A,V,f,\lambda^e,\lambda^t)$,令$X\subseteq U,R\subseteq C,a_i\in C-R,R'=R\bigcup\{a_i\}$,则属性$a_i$在属性子集$R$所构成的粒度空间中的属性代价贡献率为

$$\omega(R,a_i)=\frac{\lambda_{R(X)}-\lambda_{R'(X)}}{\lambda_{a_i}^t|\mathrm{BND}_R(X)|}. \tag{6}$$

$\omega(R,a_i)$的物理意义表示在属性集R对应的粒度空间下,引入属性a_i后,a_i引起的单位测试代价值的增长所降低的X的近似集的误分类代价值. 根据属性代价贡献率这一定义,在对问题求解时,可以根据当前粒度空间选取$\omega(R,a_i)$值最大的属性来构建更细的粒度空间. 基于这一思想,本文给出一种代价敏感的粒度寻优算法. 算法的主要思想是:首先需要初始化一个属性集$P=\{\}$. 然后,在条件属性集C中选取属性代价贡献率最大的属性细化当前粒度空间,标记被选取的属性为p_i,令$P=P\bigcup\{p_i\},C=C-\{p_i\}$,判断$C$是否为空集,如果不为空集,则继续在$C$中选取属性代价贡献率最大的属性细化粒度空间,直至$C$为空集. 将第1次选取的属性标记为$p_1$,第

2次选取的属性标记为p_2,依次类推,此时可以根据属性集P中选取属性的先后顺序构建一个层次空间结构,即$P_0 \to P_1 \to \ldots \to P_n$,其中$P_0 = \{\}, P_1 = \{p_1\}, P_n = \{p_1, p_2, \cdots, p_n\}$. 用$\mathrm{Cost}_{P_i}(X)$表示在第$i$层空间中选取属性所付出的测试代价和构建$X$的近似集的误分类代价之和. 最后,寻找$\mathrm{Cost}_{P_i}(X)$值最小所对应的粒度空间作为最优粒度空间,并在该空间下构建目标概念的误分类代价的近似集. 该算法的具体步骤如算法1所示,算法流程如图2所示.

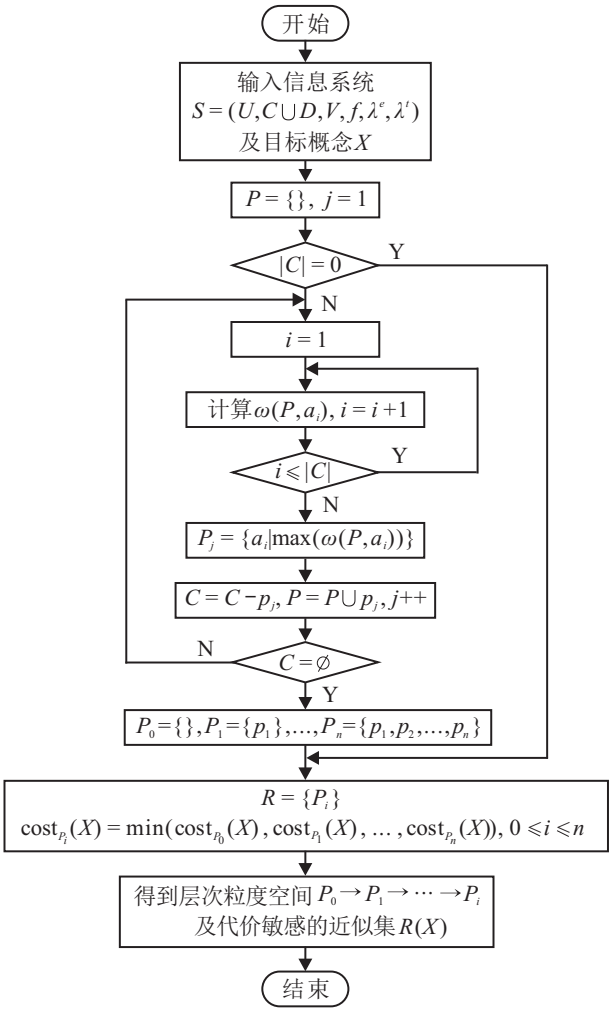


图 2　代价敏感的粒度寻优算法流程

算法1　代价敏感的粒度寻优算法.

输入: $S=(U,A,V,f,\lambda^e,\lambda^t)$及目标概念$X$;

输出: 层次粒度空间结构以及粗糙集X的近似集$R(X)$.

step 1: 初始化$P=\{\}, j=1$.

step 2: 判断C是否为空,如果为空,则执行step 6;否则执行step 3.

step 3: for $i=1:|C|$

　　计算$\omega(P,a_i)$;

　　end for

step4: $p_j = \{a_i|\max(\omega(P,a_i))\}, P = P\bigcup p_j$, $C=C-p_j, j++$.

step 5: 判断C是否为空,如果不为空,则执行step 3;如果为空,则执行step 6.

step 6: 得到属性序列$P_0=\{\}, P_1=\{p_1\}, P_2=\{p_1,p_2\},\ldots,P_n=\{p_1,p_2,\ldots,p_n\}$.

step 7: $R=\{P_i|\mathrm{Cost}_{P_i}(X)=\min(\mathrm{Cost}_{P_0}(X),\mathrm{Cost}_{P_1}(X),\ldots,\mathrm{Cost}_{P_n}(X))\}, 0\leqslant i\leqslant n$.

step 8: 得到层次粒度空间$P_1\to P_2\to\cdots\to P_i$,从而得到$R(X)$.

在计算算法的时间复杂度时,往往以最坏情况计算,即论域中的每个对象在条件属性和决策属性下得以区分. 通过对代价敏感的粒度寻优算法分析可知,算法的时间复杂度主要是由step 3、step 4以及step 7引起的. 在step 3中计算a_i的$\omega(R,a_i)$值需要求得$\lambda_{R(X)}$、$\lambda_{R'(X)}$以及$|\mathrm{BND}_R(X)|$,需要分别求属性集R以及R'下的等价划分. step 3需要循环$|C|$次,因此,step 3的时间复杂度为$O(|C|(|C|-|R|)((|R|+|R'|)|U|^2+2|U|+|\mathrm{BND}_R(X)||U|))$. 在step 4中需要找出$C-R$中$\omega(R,a_i)$值最大所对应的$a_i$,每次找最大值的时间复杂度为$O(|C-R|)$,需要循环$|C|$次,因此,step 4的时间复杂度为$O(|C||C-R|)$. 在step 7中,求得的层次粒度空间共有$|C|+1$层,需要找出总代价最小的粒度空间,此时的时间复杂度为$O(|C|+1)$. 一般情况下,对象总数$|U|$远大于属性个数$|C|$. 综上所述,代价敏感的粒度寻优算法的时间复杂度为$O(|U|^2|C|^3)$.

下面通过一个实例来说明本文算法的计算过程.

例2　给定如表2所示的测试代价和误分类代价

表 2　误分类代价的决策信息系统

U	a_1	a_2	a_3	a_4	a_5	a_6	d	λ^e
λ^t	0.1	1	2	3	4	10		
x_1	1	1	2	2.5	1	0	0	27.5
x_2	1	1	2	2.5	1	1	0	21
x_3	1	1	2	2.8	3	0	1	35
x_4	1	1	2	2.8	4	1	1	32.5
x_5	1	1	1	3	1	0	0	5.5
x_6	1	1	1	3	2	1	1	45
x_7	1	1	0	0	0	0	0	180
x_8	1	2	5	4	5	1	1	110
x_9	1	2	1	3	3	1	0	45
x_{10}	1	2	1	3	2	1	0	62.5
x_{11}	1	2	1	4	2	0	1	14
x_{12}	1	2	1	4	2	1	1	15
x_{13}	1	2	2	3	1	0	1	17.5
x_{14}	1	2	2	3	1	0	0	16
x_{15}	1	2	2	4	3	1	0	60

的决策信息系统$S=(U,A,V,f,\lambda^t,\lambda^e)$. 其中:$U=\{x_1,x_2,\ldots,x_{15}\}, C=\{a_1,a_2,\ldots,a_6\}, D=\{d\}$. 可知$U/\mathrm{IND}(D)=\{\{x_1,x_2,x_5,x_7,x_9,x_{10},x_{14},x_{15}\},\{x_3,x_4,x_6,x_8,x_{11},x_{12},x_{13}\}\}$,假设$X=\{x_3,x_4,x_6,x_8,x_{11},x_{12},x_{13}\}$,初始化$P=\{\}$,此时$U/\mathrm{IND}(P)=\{U\}$,求得$P(X)=\varnothing, \mathrm{Cost}_P(X)=269$.

为了减小总代价,需要增加属性以细化当前粒度空间. 这时要选取第1个属性,通过计算可得$\omega(P,a_1)=0,\omega(P,a_2)=0,\omega(P,a_3)=3.67,\omega(P,a_4)=3.26,\omega(P,a_5)=2.57,\omega(P,a_6)=0.09$. 根据算法1,第1次选取的属性为$a_3$,此时$P=P\bigcup\{a_3\}=\{a_3\}$,$U/\mathrm{IND}(P)=\{\{x_1,x_2,x_3,x_4,x_{13},x_{14},x_{15}\},\{x_5,x_6,x_9,x_{10},x_{11},x_{12}\},\{x_7\},\{x_8\}\}$,求得$P(X)=\{x_8\}$, $\mathrm{Cost}_P(X)=189$. 接下来选取第2个属性,通过计算可得$\omega(P,a_1)=0,\omega(P,a_2)=4.5,\omega(P,a_4)=2.51,\omega(P,a_5)=0.84,\omega(P,a_6)=0.13$. 根据算法1,第2次选取的属性为$a_2$,此时$P=P\bigcup\{a_2\}=\{a_3,a_2\}$, $U/\mathrm{IND}(P)=\{\{x_1,x_2,x_3,x_4\},\{x_5,x_6\},\{x_7\},\{x_8\},\{x_9,x_{10},x_{11},x_{12}\},\{x_{13},x_{14},x_{15}\}\}$, 求得$P(X)=\{x_1,x_2,x_3,x_4,x_5,x_6,x_8\}, \mathrm{Cost}_P(X)=143.5$.

依此进行计算,中间的计算过程不再赘述,可得后面依次选取的属性为a_4、a_5、a_1、a_6. 记得到的层次粒度空间结构为$P_0\to P_1\to P_2\to P_3\to P_4\to P_5\to P_6$. 其中:$P_0=\varnothing,P_1=\{a_3\},P_2=\{a_2,a_3\},P_3=\{a_2,a_3,a_4\},P_4=\{a_2,a_3,a_4,a_5\},P_5=\{a_1,a_2,a_3,a_4,a_5\},P_6=\{a_1,a_2,a_3,a_4,a_5,a_6\}$. 通过计算可得$\mathrm{Cost}_{P_0}(X)=269,\mathrm{Cost}_{P_1}(X)=189,\mathrm{Cost}_{P_2}(X)=143.5,\mathrm{Cost}_{P_3}(X)=103.5,\mathrm{Cost}_{P_4}(X)=114,\mathrm{Cost}_{P_5}(X)=114.2,\mathrm{Cost}_{P_6}(X)=134.2$. 由于$\min(\mathrm{Cost}_{P_0}(X),\mathrm{Cost}_{P_1}(X),\ldots,\mathrm{Cost}_{P_6}(X))=\mathrm{Cost}_{P_3}(X)$,即当粒度空间细化到由$P_3$诱导出的粒度空间时,总的代价最小. 所以$P_3$即为要寻找的最优粒度空间,且所构建出来的层次粒度空间的属性引入序列为$\{a_3\}\to\{a_3,a_2\}\to\{a_3,a_2,a_4\}$,在$P_3$下求得的误分类代价的近似集为$P_3(X)=\{x_3,x_4,x_5,x_6,x_8,x_{11},x_{12},x_{13},x_{14}\}$.

通过例2可以看出,从代价敏感的角度出发,可以根据现有信息构建出一个层次粒度空间结构,并且在这个层次粒度空间结构中可以寻找出测试代价与误分类代价之和最小的粒度空间,并求出误分类代价的粗糙集近似集.

4 实验及分析

为了验证代价敏感的粒度寻优算法的有效性,本文从UCI数据集中随机选取3个数据集进行实验,并且每个数据集在两种不同的代价环境下进行实验. 数据集的具体信息如表3所示. 实验所用计算机硬件环境为:内存为8 G,CPU为Intel(R) Core(TM)i5-4590 3.30 GHz. 软件环境为Windows10,64位操作系统,算法实现的过程使用Matlab2014平台. 其中Post-Operative Patient数据集中有3个对象个别属性值缺失,因此在实验时删除了这3个对象,并且这3个对象的删除不影响实验的进行.

表3 数据集信息表

no.	dataset	$\lvert U\rvert$	$\lvert C\rvert$	classes
1	Post-Operative Patient	90	8	3
2	Liver Disorders	345	6	2
3	Nursery	12 960	8	5

在粒度寻优过程中,不同的代价参数将诱导出不同的层次粒度空间. 因此,在对以上3个数据集进行实验时,本文将从不同的代价参数出发,寻找适应给定代价场景下的合适的粒度空间,并在该粒度空间下求出目标概念的误分类代价的近似集.

测试代价是由所获取的对象属性值而产生的,误分类代价是由原本属于某一类的对象因错误划分到另一类而产生的. 现实中,测试代价可以是金钱、时间等多种形式,例如当人们去医院就诊时,进行验血、做核磁共振都是获取属性值的过程,是需要花费一定费用的,这些费用即为测试代价;而医生在诊断过程中,会根据化验结果对病人是否患病进行判断,如果将原本没有患病的健康人误诊为患病,则原本健康的人就需要接受治疗,治疗过程所花费的费用即为误分类代价的一种体现. 现实中,可以用金钱或其他形式来衡量这两种代价参数. 由于数据集中没有给定对象的误分类代价以及属性的测试代价,在进行实验时,对于对象误分类代价参数的选取,本文采用随机生成方式,并且每个数据集将采用两组不同的代价参数进行实验. 数据集中的数据根据决策属性可以划分为不同的类别,在实验过程中,将选取数据集中的某一类数据作为目标概念.

下面将给出不同数据集进行实验时的代价参数信息. Post-Operative Patient数据集、Liver Disorders数据集、Nursery 数据集的两次实验的代价参数分别如表4、表5和表6所示. 其中,每个数据集的第1组实验每个对象的误分类代价参数均取0～50之间的随机整数值,第2组实验每个对象的误分类代价参数均取0～200之间的随机整数值. 实验的目的是寻求一个合适的粒度空间以求得给定目标概念的近似集,而整

个算法是以代价敏感为出发点,因此,在实验结果中将展示各个粒度空间下求得目标概念的误分类代价的近似集的误分类代价,得到该粒度空间的测试代价以及测试代价与误分类代价之和. 由于在实验中,目标概念中对象数量较大,在这里就不一一列举目标概念的对象以及求得的目标概念的代价敏感的近似集中的对象,实验部分将着重展示得到的层次粒度空间结构以及在各个粒度空间下,求得目标概念代价敏感的近似集的总代价、测试代价和误分类代价.

表4　Post-Operative Patient数据集代价参数

	属性	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
1	测试代价	0.5	0.5	0.6	0.9	1.5	1.1	2.5	1
	误分类代价	0～50之间的随机整数值							
2	测试代价	1	2	3	1	10	12	6	1
	误分类代价	0～200之间的随机整数值							

表5　Liver Disorders数据集代价参数

	属性	a_1	a_2	a_3	a_4	a_5	a_6
1	测试代价	4	5	3	6	5	2
	误分类代价	0～50之间的随机整数值					
2	测试代价	10	8	9	7	10	6
	误分类代价	0～200之间的随机整数值					

表6　Nursery数据集代价参数

	属性	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
1	测试代价	1	2	4	3	5	2	3	1
	误分类代价	0～50之间的随机整数值							
2	测试代价	2	5	6	4	3	8	9	1
	误分类代价	0～200之间的随机整数值							

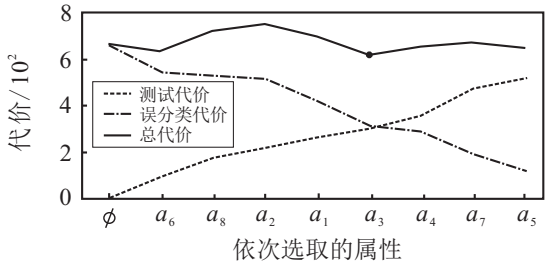
通过使用代价敏感的粒度寻优算法,将在不同代价参数环境下寻找一个合适的粒度空间,以达到测试代价与误分类代价之和最大化降低的目的. 针对Post-Operative Patient数据集,实验结果如表7和图3所示. 从实验结果可以看出,在第1组代价参数下,粒度空间由属性集$\{a_1, a_2, a_3, a_6, a_8\}$诱导得到且选取属性的顺序为$a_6 \to a_8 \to a_2 \to a_1 \to a_3$时,构建目标概念的代价敏感的近似集时总代价最小,为623.1;在第2组代数参数下,选取属性顺序为$a_8 \to a_1 \to a_4 \to a_2 \to a_3 \to a_7$时,构建目标概念的代价敏感的近似集时总代价最小,为1 435.

针对Liver Disorders数据集,实验结果如表8和图4所示. 可以看出,在第1组代价参数下,选取属性的顺序为$a_3 \to a_6 \to a_1$时,构建目标概念的代价敏感的近似集时总代价最小为2 188;在第2组代数参数下,选取属性顺序为$a_2 \to a_4 \to a_6 \to a_3$时,构建目标概念的代价敏感的近似集时总代价最小,为5 088.

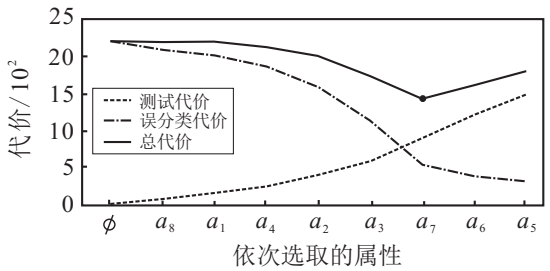
表7　Post-Operative Patient数据集下的实验结果

	选取属性顺序	∅	a_6	a_8	a_2	a_1	a_3	a_4	a_7	a_5
1	总代价	666	638.7	719.7	749.2	691.7	623.1	647.2	669.2	644.2
	测试代价	0	95.7	181.7	224.2	265.7	307.1	360.2	480.2	519.2
	误分类代价	666	543	538	525	426	316	287	189	125
	划分正确率	0.724	0.736	0.747	0.759	0.781	0.804	0.828	0.897	0.931

	选取属性顺序	∅	a_8	a_1	a_4	a_2	a_3	a_7	a_6	a_5
2	总代价	2 222	2 180	2 206	2 125	2 007	1 730	1 435	1 613	1 800
	测试代价	0	87	173	257	411	594	894	1 218	1 478
	误分类代价	2 222	2 093	2 033	1 868	1 596	1 136	541	395	322
	划分正确率	0.724	0.736	0.747	0.759	0.781	0.816	0.885	0.897	0.931



(a)　第1组代价参数下的粒度寻优



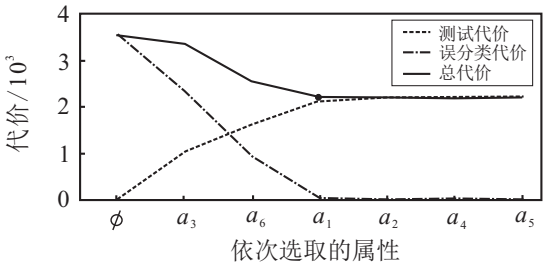
(b)　第2组代价参数下的粒度寻优

图3　Post-Operative Patient数据集下两种代价参数的粒度寻优结果

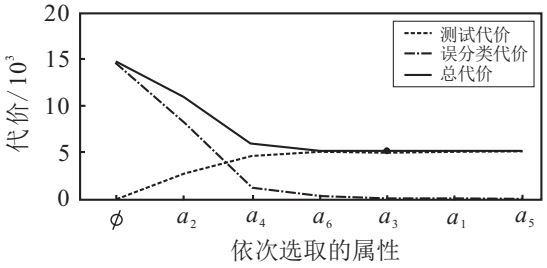
表8　Liver Disorders数据集下的实验结果

	选取属性顺序	∅	a_3	a_6	a_1	a_2	a_4	a_5
1	总代价	3 553	3 394	2 544	2 188	2 200	2 200	2 200
	测试代价	0	1 035	1 627	2135	2 200	2 200	2 200
	误分类代价	3 553	2 359	917	53	0	0	0
	划分正确率	0.580	0.675	0.870	0.982	1	1	1

	选取属性顺序	∅	a_2	a_4	a_6	a_3	a_1	a_5
2	总代价	15 031	11 008	5 927	5 191	5 088	5 088	5 088
	测试代价	0	2 760	4 797	5 043	5 088	5 088	5 088
	误分类代价	15 031	8 248	1 130	148	0	0	0
	划分正确率	0.580	0.733	0.945	0.994	1	1	1



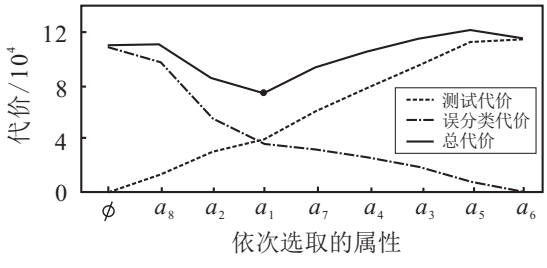
(a) 第1组代价参数下的粒度寻优



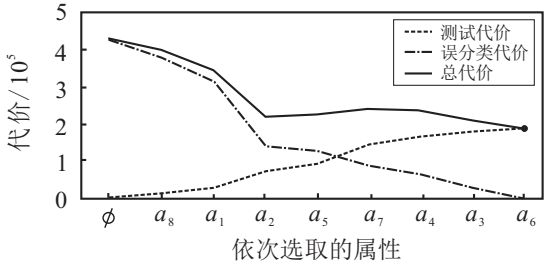
(b) 第2组代价参数下的粒度寻优

图 4 Liver Disorders 数据集下两种代价参数的粒度寻优结果

针对 Nursery 数据集,实验结果如图 5 和表 9 所示. 可以看出,在第 1 组代价参数下,选取属性的顺序为$a_8 \to a_2 \to a_1$时,构建目标概念的代价敏感的近似集时总代价最小,为 74 774;在第 2 组代数参数下,选取属性顺序为$a_8 \to a_1 \to a_2 \to a_5 \to a_7 \to a_4 \to a_3 \to a_6$时,构建目标概念的代价敏感的近似集时总代价最小,为 188 128.



(a) 第1组代价参数下的粒度寻优



(b) 第2组代价参数下的粒度寻优

图 5 Nursery 数据集下两种代价参数的粒度寻优结果

表 9 Nursery 数据集下的实验结果

	选取属性顺序	∅	a_8	a_2	a_1	a_7	a_4	a_3	a_5	a_6
1	总代价	109 982	111 129	85 191	74 774	93 314	105 161	113 916	120 869	114 820
	测试代价	0	12 960	30 240	38 880	60 480	78 336	95 232	113 052	114 820
	误分类代价	109 982	98 169	54 951	35 894	32 834	26 825	18 684	7 817	0
	划分正确率	0.671	0.710	0.832	0.892	0.901	0.919	0.942	0.966	1
	选取属性顺序	∅	a_8	a_1	a_2	a_5	a_7	a_4	a_3	a_6
2	总代价	428172	393 576	347 014	214 963	226 205	242 549	234 513	212 201	188 128
	测试代价	0	12 960	30 240	73 440	95 040	149 472	167 136	181 056	188 128
	误分类代价	428 172	380 616	316 774	141 523	131 165	93 077	67 377	31 145	0
	划分正确率	0.671	0.710	0.758	0.892	0.901	0.930	0.942	0.966	1

从以上 3 个数据集的实验结果可以看出,随着粒度空间的细化,测试代价逐步增大,而误分类代价逐步降低,并且在构建误分类代价的近似集时,划分正确率呈单调递增,这与现实生产中的情况相吻合. 而针对同一数据集和同一目标概念,当选取不同的代价参数时,所得到的层次空间不一定相同,即便是改变一个代价参数也可能引起整个层次粒度空间结构的改变,进而得到的目标概念的误分类代价的近似集可能也是不一样的. 这种以代价为诱导因子来构建合适的粒度空间也是本文算法代价敏感的体现,更加符合人类认知.

本文提出的代价敏感的粒度寻优算法是一种启发式寻优算法,根据该算法可以在给定代价场景下求得当前启发式函数上总代价最小的粒度空间,并将该粒度空间用于构建给定目标概念的误分类代价的近似集,但并不能保证所求得的粒度空间下构建目标概念的近似集所产生的测试代价与误分类代价之和是全局最优的.

5 结 论

粗糙集理论经过多年的发展,无论是在理论研究还是实际应用上都取得了许多的成果,为数据挖

掘、机器学习、模式识别等领域提供了重要的理论基础. 在前期的研究工作中,笔者提出了利用当前知识空间中的知识粒构建目标概念的近似集的方法. 但是在现实生产中,代价信息是客观存在的,在问题求解中决策者需要充分考虑代价这一因素. 因此,本文从误分类代价最小化角度出发,给出了误分类代价的粗糙集近似集模型,并讨论了该模型的相关性质. 进一步地,为了能在多粒度空间中寻找一个合适的粒度空间对目标概念进行近似描述,从而达到减小测试代价与误分类代价之和的目的,本文给出了属性代价贡献率的定义,并基于该定义,给出了一种代价敏感的粒度寻优算法. 通过该算法可以求得现有代价环境下的一个合适层次粒度空间结构,并且能在当前认知环境下选取合理的粒度空间对目标概念进行近似描述,这种渐近式求解过程符合人们的问题求解策略. 最后,利用UCI数据集对该算法进行了验证,实验结果表明了该算法的有效性和实用性. 这些研究工作进一步完善了粗糙集近似集理论,丰富了粗糙集理论,希望该研究工作能够推动不确定性人工智能的发展,扩展粗糙集理论模型及其应用.

参考文献(References)

[1] Zadeh L A. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338-353.

[2] Pawlak Z. Rough sets[J]. International Journal of Computer Information Sciences, 1982, 11(5): 341-356.

[3] Pawlak Z, Skowron A. Rudiments of rough sets[J]. Information Sciences, 2007, 177(1): 3-27.

[4] 张钹, 张铃. 问题求解的理论及应用[M]. 北京: 清华大学出版社, 2007: 1-399.
(Zhang B, Zhang L. Theory and applications of problem solving[M]. Beijing: Tsinghua University Press, 2007: 1-399.)

[5] Zhang L, Zhang B. The quotient space theory of problem solving[C]. The 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Berlin: Springer, 2003, 2639: 11-15.

[6] 李德毅, 刘常昱, 杜鹢, 等. 不确定性人工智能[J]. 软件学报, 2004, 15(11): 1583-1594.
(Li D Y, Liu C Y, Du Y, et al. Artificial intelligence with uncertainty[J]. Journal of Software, 2004, 15(11): 1583-1594.)

[7] 李德毅, 孟海军, 史雪梅. 隶属云和隶属云发生器[J]. 计算机研究与发展, 1995, 32(6): 15-20.
(Li D Y, Meng H J, Shi X M. Membership cloud and membership cloud generators[J]. Journal of Computer Research and Development, 1995, 32(6): 15-20.)

[8] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229-1246.
(Wang G Y, Yao Y Y, Yu H. A survey on rough set theory and applications[J]. Chinese Journal of Computers, 2009, 32(7): 1229-1246.)

[9] 张文修, 吴伟志. 粗糙集理论介绍和研究综述[J]. 模糊系统与数学, 2000, 14(4): 1-12.
(Zhang W X, Wu W Z. An introduction and a survey for the studies of rough set theory[J]. Fuzzy Systems and Mathematics, 2000, 14(4): 1-12.)

[10] 张清华, 胡荣德, 姚龙洋, 等. 基于属性重要度的风险决策粗糙集属性约简[J]. 控制与决策, 2016, 31(7): 1199-1205.
(Zhang Q H, Hu R D, Yao L Y, et al. Risk DTRS attribute reduction based on attribute importance[J]. Control and Decision, 2016, 31(7): 1199-1205.)

[11] 官礼和, 王国胤, 胡峰. 一种基于属性序的决策规则挖掘算法[J]. 控制与决策, 2012, 27(2): 313-316.
(Guan L H, Wang G Y, Hu F. A decision rules mining algorithm based on attribute order[J]. Control and Decision, 2012, 27(2): 313-316.)

[12] 黄恒秋, 曾玲, 黎利辉. 混合值不完备系统的双邻域粗糙集分类方法[J]. 控制与决策, 2018, 33(7): 1207-1214.
(Huang H Q, Zeng L, Li L H. Double-neighborhood rough set classification method in incomplete decision system with hybrid value[J]. Control and Decision, 2018, 33(7): 1207-1214.)

[13] Ziarko W. Probabilistic rough sets[C]. International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing. Berlin: Springer, 2005: 283-293.

[14] Didier D，Henri P. Rough fuzzy sets and fuzzy rough sets[J]. International Journal of General Systems, 1990, 17(2/3): 191-209.

[15] 于洪, 王国胤, 姚一豫. 决策粗糙集理论研究现状与展望[J]. 计算机学报, 2015, 38(8): 1628-1639.
(Yu H, Wang G Y, Yao Y Y. Current research and future perspectives on decision-theoretic rough sets[J]. Chinese Journal of Computers, 2015, 38(8): 1628-1639.)

[16] Yao Y Y. Decision-theoretic rough set models[C]. International Conference on Rough Sets and Knowledge Technology. Berlin: Springer, 2007: 1-12.

[17] Ziarko W. Variable precision rough set model[J]. Journal of Computer and System Sciences, 1993, 46(1): 39-59.

[18] Hu Q H, Yu D R, Liu J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577-3594.

[19] Yang X B, Zhang M, Dou H L, et al. Neighborhood

systems-based rough sets in incomplete information system[J]. Knowledge-Based Systems, 2011, 24(6): 858-867.

[20] 张清华, 王国胤, 肖雨. 粗糙集的近似集[J]. 软件学报, 2012, 23(7): 1745-1759.
(Zhang Q H, Wang G Y, Xiao Y. Approximation sets of rough sets[J]. Journal of Software, 2012, 23(7): 1745-1759.)

[21] 张清华, 薛玉斌, 王国胤. 粗糙集的最优近似集[J]. 软件学报, 2016, 27(2): 295-308.
(Zhang Q H, Xue Y B, Wang G Y. Optimal approximation sets of rough sets[J]. Journal of Software, 2016, 27(2): 295-308.)

[22] 姚龙洋, 张清华, 胡帅鹏, 等. 基于近似集与粒子群的粗糙熵图像分割方法[J]. 计算机科学与探索, 2016, 10(5): 699-708.
(Yao L Y, Zhang Q H, Hu S P, et al. Rough entropy for image segmentation based on approximation sets and particle swarm optimization[J]. Journal of Frontiers of Computer Science and Technology, 2016, 10(5): 699-708.)

[23] Zhang Q H, Guo Y L, Xiao Y. Attribute reduction based on approximation set of rough set[J]. Journal of Computational Information Systems, 2014, 10(16): 6859-6866.

[24] Zhang Q H, Yang J J, Yao L Y. Attribute reduction based on rough approximation set in algebra and information views[J]. IEEE Access, 2016, 4: 5399-5407.

[25] Yang Q, Wu X D. The 10 challenging problems in data mining research[J]. International Journal of Information Technology and Decision Making, 2006, 5(4): 597-604.

[26] Domingos P. MetaCost: A general method for making classifiers cost-sensitive[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle: ACM Press, 1999, 99: 155-164.

[27] 廖淑娇. 代价敏感粒计算若干方法的研究[D]. 成都: 电子科技大学信息与软件工程学院, 2018.
(Liao S J. Research on cost-sensitive granular computing approaches[D]. Chengdu: School of Information and Software Engineering, University of Electronic Science and Technology of China, 2018.)

[28] Min F, Hu Q H, Zhu W. Feature selection with test cost constraint[J]. International Journal of Approximate Reasoning, 2014, 55(1): 167-179.

[29] Min F, He H P, Qian Y H, et al. Test-cost-sensitive attribute reduction[J]. Information Sciences, 2011, 181(22): 4928-4942.

[30] Liao S J, Zhu Q X, Qian Y H, et al. Multi-granularity feature selection on cost-sensitive data with measurement errors and variable costs[J]. Knowledge-Based Systems, 2018, 158: 25-42.

[31] Yang J, Wang G Y, Zhang Q H, et al. Optimal granularity selection based on cost-sensitive sequential three-way decisions with rough fuzzy sets[J]. Knowledge-Based Systems, 2019, 163: 131-144.

[32] Yao Y Y, Wong S K M, Lingras P. A decision-theoretic rough set model[J]. Methodologies for Intelligent Systems, 1990, 5: 17-24.

[33] Yang X B, Qi Y S, Song X N, et al. Test cost sensitive multigranulation rough set: Model and minimal cost selection[J]. Information Sciences, 2013, 250: 184-199.

[34] Hunt E B, Marin J, Stone P J. Experiments in Induction[M]. New York: Academic Press, 1966: 651-653.

[35] Min F, Zhu W. Minimal cost attribute reduction through backtracking[J]. Communications in Computer and Information Sciences, 2011, 258: 100-107.

作者简介

张清华(1974–), 男, 教授, 博士生导师, 从事不确定信息处理、粗糙集与粒计算等研究, E-mail: zhangqh@cqupt.edu.cn;

刘凯旋(1992–), 男, 硕士生, 从事不确定信息处理、粗糙集与粒计算的研究, E-mail: 1025160455@qq.com;

高满(1994–), 男, 硕士生, 从事不确定信息处理、粗糙集与三支决策的研究, E-mail: gaomandaner@qq.com.

(责任编辑：李君玲)